ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ

ΤΜΗΜΑ ΕΠΙΣΤΗΜΗΣ ΥΠΟΛΟΓΙΣΤΩΝ

ΠΑΡΟΥΣΙΑΣΗ / ΕΞΕΤΑΣΗ ΜΕΤΑΠΤΥΧΙΑΚΗΣ ΕΡΓΑΣΙΑΣ

Τσέλας Χρήστος
Μεταπτυχιακός Φοιτητής

Τμήμα Επιστήμης Υπολογιστών, Πανεπιστήμιο Κρήτης
Επόπτης Μεταπτ. Εργασίας: Αναπλ. Καθηγητής, Ι. Τσαμαρδίνος

Δευτέρα, 30/10/2017, 16:00
Αίθουσα K206 ,Τμήμα Επιστήμης Υπολογιστών, Πανεπιστήμιο Κρήτης

" Η Κατασκευή Κρυμμένων Χαρακτηριστικών για Γονιδιακές Εκφράσεις Βελτιώνει την Προβλεπτική ικανότητα "

ΠΕΡΙΛΗΨΗ

Η ανάλυση γονιδιακών εκφράσεων στοχεύει στη βελτίωση της κατανόησης των ενδογενών κυτταρικών διεργασιών και συμβάλλει στην επιτυχή εφαρμογή της εξατομικευμένης ιατρικής. Η εμφάνιση των τεχνολογιών γονιδιακών εκφράσεων υψηλών αποδόσεων, όπως μικροσυστοιχίες (microarrays) και αλληλουχία RNA (RNAseq) καθώς και η πρόσφατη μείωση του κόστους, οδήγησαν στην έκρηξη δημόσιων-διαθέσιμων συνόλων δεδομένων. Τα παραγόμενα σύνολα δεδομένων είναι αναπόφευκτα μεγάλης διάστασης με τυπικά μικρό μέγεθος δείγματος που περιορίζει σοβαρά τη δυνατότητα δημιουργίας αναπαραγώγιμων προγνωστικών μοντέλων. Η δυνατότητα αύξησης της προβλεπτικής ισχύος χωρίς απώλεια πληροφοριών του μετρηθέντος γονιδιώματος σε ένα νεοσύστατο σύνολο δεδομένων είναι ύψιστης σημασίας. Παρά το γεγονός ότι διάφορες μελέτες έχουν προσπαθήσει να επιτύχουν μείωση των διαστάσεων και συγχώνευση συνόλων δεδομένων, ώστε να αυξηθεί η απόδοση και η ευρωστία της ταξινόμησης, εξακολουθούν να υπάρχουν προκλήσεις κυρίως λόγω του περιορισμένου αριθμού δεδομένων καθώς και της τεχνολογικής ποικιλομορφίας και ετερογένειας στα σύνολα δεδομένων.

Αξιοποιώντας την πλεοναστικότητα των γονιδιακών δεδομένων, κατασκευάσαμε καθολικούς κρυμμένους χώρους μικρότερων διαστάσεων του γονιδιώματος, χρησιμοποιώντας διάφορες προσεγγίσεις μείωσης των διαστάσεων και ένα ποικίλο σύνολο συνόλων δεδομένων. Οι τεχνικές Principal Component Analysis (PCA), kernel PCA και Neural Network Autoencoder εφαρμόστηκαν σε σύνολα δεδομένων από τέσσερις διαφορετικές πλατφόρμες. Ενώ οι γραμμικές τεχνικές έδειξαν καλύτερες επιδόσεις ανασυγκρότησης, οι μη γραμμικές προσεγγίσεις ήταν σε θέση να καταγράψουν πιο πολύπλοκες γονιδιακές αλληλεπιδράσεις, απολαμβάνοντας έτσι ισχυρότερη προβλεπτική δύναμη. Όταν νεοφανή σύνολα γονιδιακών εκφράσεων προβάλλονται σε ένα κρυμμένο χώρο 200 διαστάσεων, η προβλεπτική ισχύς βελτιώθηκε. Επιπλέον, πραγματοποιήσαμε ένα πείραμα μεγάλης κλίμακας, όπου οι μέθοδοι μείωσης των διαστάσεων εκπαιδεύτηκαν σε ένα σύνολο 59864 μοναδικών δειγμάτων. Η ισχύς ταξινόμησης βελτιώθηκε περαιτέρω ειδικά για την τεχνική Autoencoder. Απροσδόκητα, η στατιστική μεταβλητότητα των πρόσθετων συνόλων δεδομένων αύξησε την απόδοση ταξινόμησης υπονοώντας ότι μαθεύτηκαν καλύτερα περίπλοκα βιολογικά χαρακτηριστικά. Επιπλέον, εξετάσαμε τη δυνατότητα αύξησης των δεδομένων χρησιμοποιώντας δεδομένα από διάφορες πλατφόρμες, κατασκευάζοντας ένα ενδιάμεσο χώρο χαρακτηριστικών που δείχνει ότι όταν οι πλατφόρμες μοιράζονται κοινά χαρακτηριστικά (όπως GLP570 και GLP96) βελτιώνεται η προβλεπτική απόδοση.

Tselas Christos
M.Sc. Thesis

Computer Science Department
University of Crete
Master's Thesis Supervisor: Associate Professor, I. Tsamardinos

Monday, 30/10/2017, 16:00
Room K206, Computer Science Dept., University of Crete

"Latent Feature Construction for Gene Expressions Improves Predictions"

ABSTRACT

Gene expression analysis aims to improve the understanding of the intrinsic cellular processes and contribute towards the successful implementation of personalized medicine. The advent of high-throughput gene expression technologies such as microarrays and RNA-sequencing (RNAseq) as well as the recent reduction of cost resulted in an explosion of publicly-available datasets. The generated datasets are inevitably high-dimensional with typically small sample size that severely limits the potential for developing reproducible prognostic models. Being able to

increase the predictive power without losing the information of the measured genome on a newly-produced dataset is of paramount importance. Despite the fact that various studies attempt to perform dimensionality reduction and dataset integration so as to increase classification performance and robustness, there are still challenging issues primarily due to the limited number of data as well as the technological diversity and heterogeneity across the datasets.

Exploiting the redundancy of genomics data, we constructed low-dimensional, universal, latent feature spaces of the genome utilizing several dimensionality reduction approaches and a diverse set of curated datasets. Standard Principal Component Analysis (PCA), kernel PCA and Neural Network Autoencoders were applied on datasets from four different platforms. While linear techniques showed better reconstruction performance, nonlinear approaches were able to capture more complex gene interactions, and thus enjoyed stronger classification power. When newly-seen gene expression datasets projected to a latent space of 200 dimensions, the classification power was improved. Moreover, we performed a large-scale experiment where the dimensionality reduction methods were trained on an integrated set of 59864 unique samples. The classification power was further improved especially for Autoencoder. Rather surprisingly, the statistical variability of the additional datasets increased the classification performance implying that intricate biological features were better learn. We additionally tested the possibility of cross-platform data augmentation by constructing an intermediate feature space showing that when platforms share common characteristics (such as GLP570 and GLP96) the predictive performance was also improved.